Network Model-Assisted Inference from Respondent-Driven Sampling Data

August 2, 2011

Abstract

Respondent-Driven Sampling is a method to sample hard-to-reach human populations by link-tracing over their social networks. Beginning with a convenience sample, each person sampled is given a small number of uniquely identified coupons to distribute to other members of the target population, making them eligible for enrollment in the study. This can be an effective means to collect large diverse samples from many populations.

Inference from such data requires specialized techniques for two reasons. Unlike in standard sampling designs, the sampling process is both partially beyond the control of the researcher, and partially implicitly defined. Therefore, it is not generally possible to directly compute the sampling weights necessary for traditional design-based inference. Any likelihood-based inference requires the modeling of the complex sampling process often beginning with a convenience sample. We introduce a model-assisted approach, resulting in a design-based estimator leveraging a working model for the structure of the population over which sampling is conducted.

We demonstrate that the new estimator has improved performance compared to existing estimators and is able to adjust for the bias induced by the selection of the initial sample. We present sensitivity analyses for unknown population sizes and the misspecification of the working network model. We develop a bootstrap procedure to compute measures of uncertainty. We apply the method to the estimation of HIV prevalence in a population of injecting drug users (IDU) in the Ukraine, and show how it can be extended to include application-specific information.

Keywords: Hard-to-reach population sampling; Link-tracing; Network sampling; Social networks; Exponential-family random graph model

1 Introduction

There is much interest in estimating features of hard-to-reach human populations. Such populations are characterized by the lack of a serviceable population sampling frame. In some settings, the target population is well-connected by a network of social relations. *Link-tracing* sampling strategies such as *snowball sampling* (Goodman, 1961) and *respondent-driven sampling* (RDS) (Heckathorn, 1997) are often used to leverage those social relations to sample beyond the small subgroup available to researchers. In these settings, subsequent samples are identified and selected based on their social ties with other members of the target population. The statistical literature dealing with such strategies (Frank, 1971; Goodman, 1961; Thompson, 1990; Thompson and Frank, 2000), typically assumes an idealized setting in which the initial sample is assumed to be a probability sample from the target population. The applied literature on the other hand Trow (1957); Watters and Biernacki (1989), has traditionally recognized that this is impractical, and therefore treated link-tracing samples (typically referred to as snowball samples, despite Goodman's probabilistic framing) as convenience samples for which probability-based inferential methods are unfounded.

The work of Heckathorn and colleagues (Heckathorn, 1997, 2007; Salganik and Heckathorn, 2004; Volz and Heckathorn, 2008) around the RDS specialization of link-tracing sampling is innovative in reducing the number of links followed per respondent, such that many waves of sampling are fostered, decreasing the dependence of the final sample on the initial convenience sample. The second main innovation of the RDS paradigm is in the *respondent-driven* nature of the sampling process in which subsequent samples are selected by the passing of coupons by current sample members, thus reducing the confidentiality concerns often present in hard-to-reach marginalized populations. While this approach does reduce the dependence of the final sample on the initial sample, it is possible for substantial bias to remain based on the initial sample of seeds, as studied

in simulations by Gile and Handcock (2010) and illustrated empirically by Johnston (2010). Current estimation methods (Gile, 2011; Heckathorn, 1997, 2007; Salganik and Heckathorn, 2004; Volz and Heckathorn, 2008), however, do not correct for biases introduced by seed selection. A common feature of networked populations is that the social ties are often more likely to occur between people that have similar attributes than those who do not, a tendency called *homophily* by attributes (Freeman, 1996; Lazarsfeld and Merton, 1954; McPherson, Smith-Lovin and Cook, 2001). In this paper we present a novel approach and inferential frame to correct for bias introduced by seed selection and for the effects of homophily. In particular, we treat the problem of estimation of the population proportion of a binary covariate in populations where there exists homophily on the covariate of interest, based on a branching link-tracing sample beginning with seeds selected with bias with respect to that covariate.

There is a varied formal statistical literature on inference from link-tracing network samples. All of this work, however, involves the assumption that the initial sample is a probability sample drawn from a well-defined sampling frame, and that subsequent sampling is *adaptive*, or dependent on population characteristics only through their observed portions (Thompson and Seber, 1996). In the design-based framework, these works consider cases where sampling probabilities are known for all units in the analysis (Frank, 1971, 2005; Goodman, 1961; Thompson, 1992, 2006). Inference is then made without reference to any superpopulation model. In the likelihood frame, the literature treats cases where the adaptive sampling process is *amenable* to the model, and therefore the modeling can be conducted without explicit treatment of the sampling process (Handcock and Gile, 2010; Pattison, Robins, Daraganova, Wang, Koskinen and Snijders, 2009; Thompson and Frank, 2000). The traditional approach to RDS, originally due to Heckathorn (1997), represents an alternative to this paradigm. The assumption of the original probability sample is replaced by an assumption of sufficient waves of sampling to adequately reduce the dependence of the sample on the original sample.

In this paper, we concern ourselves with a case in which none of these approaches suffice. The sampling probabilities of the units are not known, making the traditional design-based approaches inadequate. The initial sample is not a probability sample, so the sample is not adaptive or amenable, and any likelihood inference must consider the sampling process as well as the population model. Such a joint modeling approach has been con-

ducted in a few works (Felix-Medina and Monjardin, 2006; Felix-Medina and Thompson, 2004; Frank and Snijders, 1994), but each of these requires an initial probability sample from some frame to allow for modeling of the sampling process. And while in some cases, the waves of sampling may be sufficient to suitably reduce the dependence on the initial sample, this is often not the case (Gile and Handcock, 2010), and we are interested in the cases when this does not hold.

We begin in Section 2 by introducing respondent-driven sampling. In Section 3, we then present our Model-Assisted inferential approach. Section 4 presents a simulation study illustrating the removal of bias introduced by the initial convenience sample. Our application to HIV prevalence estimation among injecting drug users in the Ukraine can be found in Section 5, and Section 6 presents a discussion and concluding remarks.

2 Respondent-Driven Sampling

2.1 Notation

We assume the target population consists of N people (nodes) with labels $1, \ldots, N$. Let the N-vector $\mathbf{z}$, represent a binary nodal outcome variable of interest. We refer to this variable as "infection status", such that

$$\mathbf{z}_i = \begin{cases} 0 & i \text{ not infected} \\ 1 & i \text{ infected.} \end{cases} \qquad i \in 1 \ldots N$$

We assume the target population is connected by a network of mutual relations with $N \times N$ adjacency matrix $\mathbf{y}$:

$$\mathbf{y}_{ij} = \mathbf{y}_{ji} = \begin{cases} 1 & i \text{ and } j \text{ connected} \\ 0 & i \text{ and } j \text{ not connected,} \end{cases}$$

and that this network forms a single connected component. Denote by $\mathbf{d}_i = \sum_j \mathbf{y}_{ij}$ the nodal *degree*, or number of network ties or *alters* of node i. Let $\mathbf{d} = \{\mathbf{d}_1, \ldots, \mathbf{d}_N\}$. Denote by $\mathbf{x}_i = \sum_j \mathbf{z}_j \mathbf{y}_{ij}$ the number of network ties node i shares with infected nodes, and let $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$.

2.2 Sampling Procedure

We consider an RDS procedure of the following form:

0. An small initial sample is selected from the population members accessible to researchers, typically using a convenience mechanism. They are typically 3-12 in number. They are called the *seeds* and comprise *wave* $k = 0$ of the sample.

1. Each member of wave k is given a small number (typically 2-3) of uniquely identified coupons to distribute among their alters.

2. Coupon recipients returning their coupons to the study center are subsequently enrolled in the study. A person recruited in a prior wave can not be recruited again. The wave number of a respondent is one more than that of their recruiter.

3. Steps (1) and (2) are repeated until the desired sample size is attained.

This process has proved effective at recruiting large and diverse samples from many hard-to-reach populations (Abdul-Quader, Heckathorn, McKnight, Bramson, Nemeth, Sabin, Gallagher and Jarlais, 2006), and has been widely used. It has been heavily used in the monitoring of disease prevalence and risk behaviors among high-risk populations such as sex workers, men who have sex with men, and injecting drug users (Malekinejad, Johnston, Kendall, Kerr, Rifkin and Rutherford, 2008), largely in the service of the reporting requirements of UNAIDS for all countries with concentrated HIV epidemics (UNAIDS, 2008). It is also used by the US Centers for Disease Control and Prevention in the behavioral monitoring of injecting drug users in 25 large US cities (Lansky, Abdul-Quader, Cribbin, Hall, Finlayson, Garfein, Lin and Sullivan, 2007), and has also been used in other populations such as unregulated workers (Bernhardt, Heckathorn, Milkman and Theodore, 2006) and jazz musicians (Heckathorn and Jeffri, 2001).

We represent the full sampling mechanism by the random variables:

$$\mathbf{S}_i^k = \begin{cases} 1 & \text{person } i \text{ is sampled in wave } k \\ 0 & \text{otherwise} \end{cases} \qquad i \in 1 \ldots N, k \in 0, \ldots .$$

$$\mathbf{S}_i = \sum_{k=0}^{\infty} \mathbf{S}_i^k = \begin{cases} 1 & \text{person } i \text{ is sampled} \\ 0 & \text{person } i \text{ is not sampled} \end{cases} \qquad i \in 1 \ldots N,$$

and let $\mathbf{s}^k$ denote the observed sampling vector corresponding to the people sampled in wave k. Based on the sampling procedure, we exactly observe the elements of $\mathbf{z}$, $\mathbf{d}$ and $\mathbf{x}$ corresponding to $i : \mathbf{s}_i = 1$. A variant when $\mathbf{x}$ cannot be observed directly, as in the application in Section 5, substitutes an estimate of $\mathbf{x}$ based on observed referral patterns.

Further we assume each respondent distributes a number of coupons completely at random from among their alters, with the number determined by a common distribution.

2.3 Design-based Inferential Approach

We consider design-based estimators for the population mean $\mu = \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i$. Because the sampling probabilities of the people selected through RDS are almost never explicitly known, we follow Volz and Heckathorn (2008), and Gile (2011) in constructing a model for the sampling process, and estimating sampling probabilities accordingly. We use a generalized Horvitz-Thompson estimator of the form:

$$\hat{\mu} = \frac{\sum_{i=1}^{N} \frac{\mathbf{S}_i \mathbf{z}_i}{\hat{\pi}_i}}{\sum_{i=1}^{N} \frac{\mathbf{S}_i}{\hat{\pi}_i}}, \tag{1}$$

where estimated sampling probabilities $\hat{\pi}_i = \mathbb{E}(\mathbf{S}_i|\mathscr{S})$ are computed under an approximation $\mathscr{S}$ to the true RDS sampling process. If the inclusion probabilities were known this estimator is referred to as the Häjek estimator (Lumley, 2010), and typically performs better than the corresponding Horvitz-Thompson estimator (Särndal, Swensson and Wretman, 1992). The estimators introduced by Volz and Heckathorn (2008) and Gile (2011) differ, and ours further differs, in their specification of the sampling process $\mathscr{S}$.

Most inference from RDS data approximates the sampling process as a with-replacement random walk on the space of graph nodes, with transitions along the edges or social relations. For the purpose of inference, sampling is treated as a Markov chain at equilibrium (Salganik and Heckathorn, 2004; Volz and Heckathorn, 2008). Such inference involves sampling weights proportional to the self-reported degrees which are the equilibrium sampling probabilities of the with-replacement random walk on a connected network. While this is a useful first approximation, it has several limitations. First, as highlighted in Gile (2011), this type of inference does not respect the without-replacement nature of the sampling process, which

can lead to biased estimates. Gile (2011) presents an approach correcting for this feature by substituting a without-replacement successive sampling approximation to the sampling process. Neither this, nor earlier estimators, however, address the fundamental issue of bias induced by the selection of the initial sample. Such bias is illustrated in Gile and Handcock (2010), as well as in the current paper, and correction for it is a key contribution of the present paper.

As with these earlier approaches, the first requirement of our sampling model is that it account for the different sampling probabilities by nodal degree. Unlike these other approaches, we further require our approach to account for the bias introduced by the selection of seeds in the presence of network homophily in the underlying population. This requires consideration of features of the social network, $\mathbf{y}$, in particular the homophily of the relations.

We make no assumptions about the mechanism for selecting the initial sample and will condition on the seed characteristics throughout the analysis.

If the network $\mathbf{y}$ were fully known, we could use simulation to estimate the sampling probability $\hat{\pi}_{i,\mathbf{y}} = \mathbb{E}(\mathbf{S}_i|\mathscr{S}, \mathbf{y}, \mathbf{s}^0)$ of each node, conditional on the selection of seeds, $\mathbf{s}^0$. Explicitly, we would repeatedly simulate RDS under sampling model $\mathscr{S}$ starting from $\mathbf{s}^0$ each time and compute the $\hat{\pi}_{i,\mathbf{y}}$ as the proportion of simulated samples containing node i. These could be used in (1) to form an estimator. Unfortunately, $\mathbf{y}$ is typically only partially known, and so we apply a model-assisted approach.

3 A Model-Assisted Approach

Our approach is an extension of the model-assisted design-based approaches presented in Särndal et al. (1992). Existing work in this area uses a working model form to construct estimators that are (approximately) design-unbiased, whether the model holds or not, and have smaller design variance if the model does hold. Our case is slightly different. The sampling process we consider is only locally defined, and originates at a sample with unknown distribution. We therefore cannot guarantee design-unbiasedness. In fact, we require reference to a model form to recover approximate design-unbiasedness, rather than to improve efficiency. This is because the impact of the seed characteristics on the subsequent sample is mediated by the structure of the underlying social network.

Our approach is to assume a working superpopulation model from which the network was drawn and use it to estimate sampling probabilities conditional on the selection of the initial sample.

3.1 Network Working Model

We consider models of *exponential-family random graph model* (ERGM) form (Frank and Strauss, 1986; Hunter, Goodreau and Handcock, 2008; Hunter and Handcock, 2006), conditional on the set of nodal degrees and infection statuses, and including a single additional parameter representing homophily on $\mathbf{z}$. In particular:

$$P(\mathbf{Y} = \mathbf{y}|\mathbf{z}, \mathbf{d}, \eta) = \frac{\exp(\eta g(\mathbf{y}, \mathbf{z}))}{c(\eta|\mathbf{z}, \mathbf{d})}, \tag{2}$$

where $g(\mathbf{y}, \mathbf{z}) = \sum_{i,j=1}^{N} \mathbf{y}_{ij}\mathbf{z}_i(1-\mathbf{z}_j)$, and $c(\eta|\mathbf{z}, \mathbf{d}) = \sum_{\mathbf{u}\in\mathscr{Y}(\mathbf{z},\mathbf{d})} \exp(\eta g(\mathbf{u}, \mathbf{z}))$, and the space $\mathscr{Y}(\mathbf{z}, \mathbf{d})$ consists of all binary undirected networks consistent with $\mathbf{d}$ and $\mathbf{z}$ (the dependence on $\mathbf{d}$ and $\mathbf{z}$ is suppressed below). Note that this model form, as well as the simulation procedure to follow, requires knowledge of the population size N.

Given this model form, we use the estimator (1) based on sampling weights assumed constant over equivalence classes by degree and infection status and estimated under the model:

$$\hat{\pi}_{i,\eta} = \mathbb{E}(\mathbf{S}_i|\mathscr{S}, \mathbf{z}, \mathbf{d}, \eta, \mathbf{s}^0) = \sum_{\mathbf{y}\in\mathscr{Y}(\mathbf{z},\mathbf{d})} \hat{\pi}_{i,\mathbf{y}} P(\mathbf{Y} = \mathbf{y}|\mathbf{z}, \mathbf{d}, \eta),$$

where $\hat{\pi}_{i,\mathbf{y}} = \mathbb{E}(\mathbf{S}_i|\mathscr{S}, \mathbf{y}, \mathbf{s}^0)$, as defined in Section 2.3. Note that to treat these equivalence classes, we condition on the equivalence classes of the seed nodes selected, rather than the unique identities of those nodes.

We also do not know the network working model parameter η, and therefore must estimate it from the available data. The estimator is then computed using sampling probabilities based on the estimated network working model given by $\hat{\eta}$:

$$\hat{\pi}_{i,\hat{\eta}} = \mathbb{E}(\mathbf{S}_i|\mathscr{S}, \mathbf{z}, \mathbf{d}, \hat{\eta}, \mathbf{s}^0) = \sum_{\mathbf{y}\in\mathscr{Y}(\mathbf{z},\mathbf{d})} \pi_{i,\mathbf{y}} P(\mathbf{Y} = \mathbf{y}|\mathbf{z}, \mathbf{d}, \hat{\eta}) \tag{3}$$

These are the estimated probabilities used in our proposed estimator. This requires fitting a network working model to data sampled through RDS, which we address in the next section.

3.2 Fitting the Network Working Model

Thompson and Frank (2000) and Handcock and Gile (2010) provide an approach to fitting models of form similar to (2) to data sampled through link-tracing samples. Unfortunately, these approaches require a sample that is *amenable* to the model in question. That is:

$$P(\mathbf{S}|\mathbf{y}, \mathbf{d}, \mathbf{z}) = P(\mathbf{S}|\mathbf{y}_{obs}, \mathbf{d}_{obs}, \mathbf{z}_{obs}), \tag{4}$$

where $*_{obs}$ represents the observed part of $*$, and also that the sampling and model parameters are separable. This is equivalent to the conditions for *ignorability* according to Rubin (1976) and Little and Rubin (2002). Unfortunately, in the case of RDS, condition (4) is violated by the convenience sample of seeds, which may well depend on unobserved characteristics.

Therefore, we require a novel approach to model fitting. As $\mathbf{d}$ and $\mathbf{z}$ are unknown, we construct design-based estimators of them from estimates of the sampling probabilities $\hat{\pi}_i$. Specifically, let $\mathbb{N}_{kl}$ be the number of nodes of degree k and infection status l, $k \in \{1, \ldots, N-1\}$, $l \in \{0, 1\}$ and $\mathbb{N} = \{\mathbb{N}_{kl}\}_{k=1;l=0}^{k=N-1;l=1}$. We estimate $\mathbb{N}$ and $g(\mathbf{y}, \mathbf{z})$ by,

$$\tilde{\mathbb{N}}_{kl} = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{S}_i \mathbb{I}(\mathbf{d}_i = k, \mathbf{z}_i = l)}{\hat{\pi}_i} \tag{5}$$

$$\tilde{g}(\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{N} \frac{\mathbf{S}_i \left(\mathbf{x}_i(1 - \mathbf{z}_i) + (\mathbf{d}_i - \mathbf{x}_i)\mathbf{z}_i\right)}{2\hat{\pi}_i} \tag{6}$$

where $\mathbb{I}(*)$ is the indicator function on $*$, and $\hat{\pi}_i$ is assumed constant for all $i : \mathbf{d}_i = k, \mathbf{z}_i = l$. Note that this requires the observation of $\{\mathbf{x}_i : \mathbf{S}_i = 1, i = 1, \ldots, N\}$. We then estimate η as the natural parameter corresponding to the mean value parameter $\tilde{g}(\mathbf{y}, \mathbf{x})$ with the joint degree and infection status sequence implied by $\tilde{\mathbb{N}}$. Details of this computation are given in the Supplemental Materials.

3.3 Algorithm

Note that the value of the network working model parameter, required to estimate π, in turn, depends on the value of π. We therefore apply an approach similar to self-consistency (Lee and Meng, 2007) to find a joint

solution to (5) and (6), as well as to the equations:

$$\hat{\pi}_{i,\hat{\eta}} = \mathbb{E}\left(\mathbf{S}_i | \mathscr{S}, \mathbf{z}, \mathbf{d}, \hat{\eta}(\tilde{\mathbb{N}}, \tilde{g}(\mathbf{y}, \mathbf{x}))\right) \qquad i = 1, \ldots, N. \tag{7}$$

This approach iterates between estimating the network working model parameter given values for the sampling probabilities, and then estimating the sampling probabilities given the network working model parameter. Explicitly, it is:

- Estimate $\hat{\pi}_{\hat{\eta}}$ proportional to degree d_i.
- Iterate the following steps:
  - Compute design-based estimates of statistics $\tilde{\mathbb{N}}_{kl}$ and $\tilde{g}(\mathbf{y}, \mathbf{x})$ using $\hat{\pi}_{\hat{\eta}}$ in (5) and (6).
  - Determine the working ERGM parameter η corresponding to $\tilde{\mathbb{N}}$ and $\tilde{g}(\mathbf{y}, \mathbf{x})$.
  - Simulate M networks according to the working ERG model. Estimate $\hat{\pi}_{\hat{\eta}}$ by simulated RDS sampling from the resulting networks.
- Use the resulting estimated probabilities, $\hat{\pi}_{\hat{\eta}}$, to form the weighted estimator of the quantity of primary interest:

$$\hat{\mu}_{MA} = \frac{\sum_{i=1}^{N} \frac{\mathbf{S}_i \mathbf{z}_i}{\hat{\pi}_{\hat{\eta}}}}{\sum_{i=1}^{N} \frac{\mathbf{S}_i}{\hat{\pi}_{\hat{\eta}}}}. \tag{8}$$

The iterative nature of this procedure is similar to that used for the successive sampling estimator of Gile (2011). This algorithm differs in the core process of estimating the inclusion probabilities. More details of this procedure are provided in the supplemental materials.

The simulation procedure implicit in this estimation algorithm lends itself to a realistic bootstrap approach to standard error estimation. We present such a bootstrap in the supplemental materials, along with a simulation study illustrating its performance under a variety of conditions.

4 Comparing the Model-Assisted to Existing Estimators: A Simulation Study

Gile and Handcock (2010) present an extensive simulation study of RDS based, where possible, on a set of realistic characteristics of data from the CDC pilot study of RDS (Abdul-Quader et al., 2006). For comparison purposes, our simulation study uses the same simulated populations as Gile (2011), along with extensions necessary for our sensitivity analyses.

4.1 Study Design

Our simulation study is designed around three levels of simulation:

- The generation of random networks according to specified network features
- The generation of simulated RDS samples from each network
- The estimation of infection prevalence from each set of simulated sample data.

We use variants of the network and sampling parameters to study the behavior of the proposed estimator. Descriptions and levels of these parameters are listed in Tables 1 and 2.

Table 1: Parameters of simulated networks. Default parameters given in boldface.

Parameter	Meaning	Values
Number of nodes		**1000**, 715
Prevalence	$\mu = \frac{1}{N}\sum_i \mathbf{z}_i$	**0.20**
Mean degree	$\bar{\mathbf{d}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(d_i) = \frac{1}{2}\sum_{i,j=1}^{N}\mathbb{E}(Y_{ij})$	**7**
Homophily	$R = \frac{\frac{2}{N^1(N^1-1)}\sum_{i,j}\mathbf{z}_i\mathbf{z}_j\mathbb{E}(y_{ij})}{\frac{1}{N^1N^0}\sum_{i,j}\mathbf{z}_i(1-\mathbf{z}_j)\mathbb{E}(y_{ij})}$ where $N^0 = N(1-\mu), N^1 = N\mu$	**5**, 3, 1
Activity Ratio	$w = \frac{\bar{d}^1}{\bar{d}^0} = \frac{\frac{1}{N^1}\sum_{i,j}\mathbf{z}_i\mathbb{E}(y_{ij})}{\frac{1}{N^0}\sum_{i,j}(1-\mathbf{z}_i)\mathbb{E}(y_{ij})}$	**1**, 1.8

To allow for comparability across simulation conditions, throughout our simulations, we maintain the same true recoverable prevalence, $\mu =$

0.20, the same sample size $n = 500$, and the same mean degree $\bar{\mathbf{d}} = 7$. We consider variations on the population size (hence the sample fraction), the degree of clustering or *homophily* on infection status, and differential rates of tie formation by infection status (or *activity ratio*). The parameter levels considered are summarized in Table 1.

Under each set of network parameters, networks are simulated according to an ERGM with sufficient statistics:

$$\begin{aligned} g_1(\mathbf{y}) &= \sum_{i=1}^{N} \sum_{j<i} \mathbf{y}_{ij} \mathbf{z}_i \mathbf{z}_j \\ g_2(\mathbf{y}) &= \sum_{i=1}^{N} \sum_{j<i} \mathbf{y}_{ij} (1 - \mathbf{z}_i)(1 - \mathbf{z}_j) \\ g_3(\mathbf{y}) &= \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{y}_{ij} \mathbf{z}_i (1 - \mathbf{z}_j). \end{aligned} \tag{9}$$

These three terms correspond to the unique cells of the 2×2 mixing matrix on $\mathbf{z}$, and for a given number of nodes N and prevalence μ, are uniquely defined by $\bar{\mathbf{d}}$, R, and w. Note that this model is similar to (2), but not identical. While (2) conditions on the fixed degree of each node, this model allows for stochastic variability in degrees around mean value parameters given by (9).

From each simulated network, a single RDS sample is drawn according to parameters in Table 2. A fixed number n^0 of seed nodes are selected with probability proportional to degree (the best case for the earlier estimators), from either the full population or from the infected nodes only (to simulate extreme seed bias). The simulated process treats the case of two coupons distributed by each respondent completely at random among its previously un-sampled alters. Two coupons are chosen for simplicity, and because it represents the sampling process better than either 3 (equating to the return of all coupons in practice) or 1 (resulting in non-branching chains).

For each simulation case, we simulate 1000 networks with one RDS sample from each, and we compare five estimators, as summarized in Table 3.

Table 2: Parameters of simulated RDS sampling. Default parameters given in boldface.

Parameter	Meaning	Values
Number of Seeds	$n^0 = \sum_i \mathbf{S}_i^0$	**10**, 6, 20
Seed Selection	Sequentially with probability proportional to degree from either:	**full population**, infected nodes
Branching	From each sampled node, up to n_{cup} previously unselected alters are selected completely at random for subsequent sampling. n_{cup} are selected whenever available.	**2**
Sample Size	Sampling stops when n nodes have been sampled.	**500**

Table 3: Five estimators compared in the simulation study.

Abbreviation	Source	Estimator
Mean	Naive sample mean of $\mathbf{z}_i$	$\hat{\mu}$
SH	Salganik and Heckathorn (2004)	$\hat{\mu}_{SH}$
VH	Volz and Heckathorn (2008)	$\hat{\mu}_{\mathrm{VH}}$
SS	Gile (2011)	$\hat{\mu}_{SS}$
MA	Current paper	$\hat{\mu}_{MA}$

4.2 Primary Results

We begin by studying the performance of the proposed estimator in settings where previous estimators have been found to perform well. In the first part of Figure 1, there is a relatively small sample fraction (50%, $N = 1000$), no homophily on infection status ($R = 1$), the ratio of mean degrees by infection (w) is 1, and seeds are chosen from the full population, so there is no bias induced by seed selection. In this case, none of the estimators considered exhibit bias, and the naive sample mean exhibits the lowest variance, although the variability is similar across estimators.

The second part of Figure 1 illustrates the case $\hat{\mu}_{SS}$ is designed to address. In this case, the sample fraction is large (about 70%, $N = 715$), and infected nodes have mean degree 80% higher than that of uninfected ($w = 1.8$). In this case there is still no homophily ($R = 1$), and no seed bias. Here, the higher-degree infected nodes are over-represented in the sample, resulting in positive bias in the sample mean. Because of assumed linear mapping from degree to sampling probability, $\hat{\mu}_{SH}$ and $\hat{\mu}_{\mathrm{VH}}$ overcorrect for this feature, resulting in negative bias. $\hat{\mu}_{SS}$ and $\hat{\mu}_{MA}$ appropri-

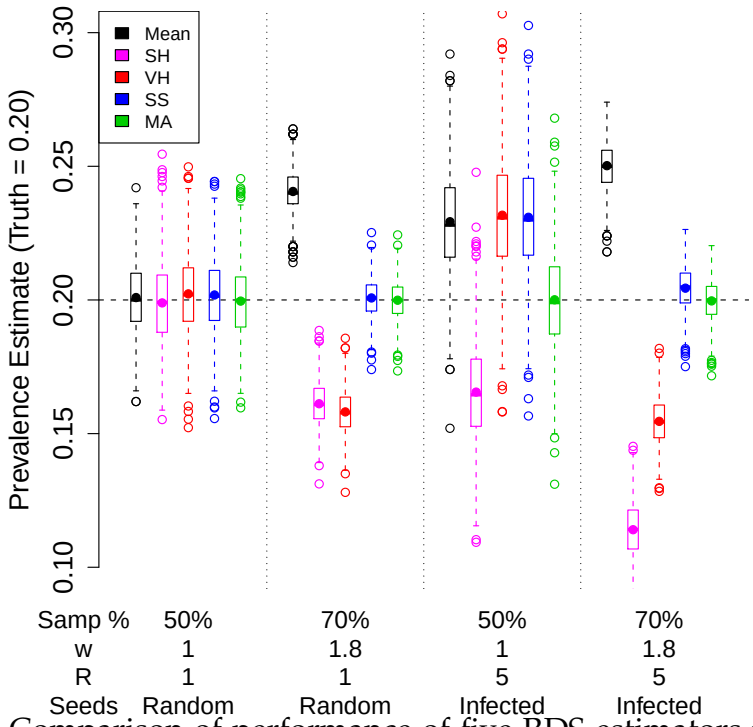


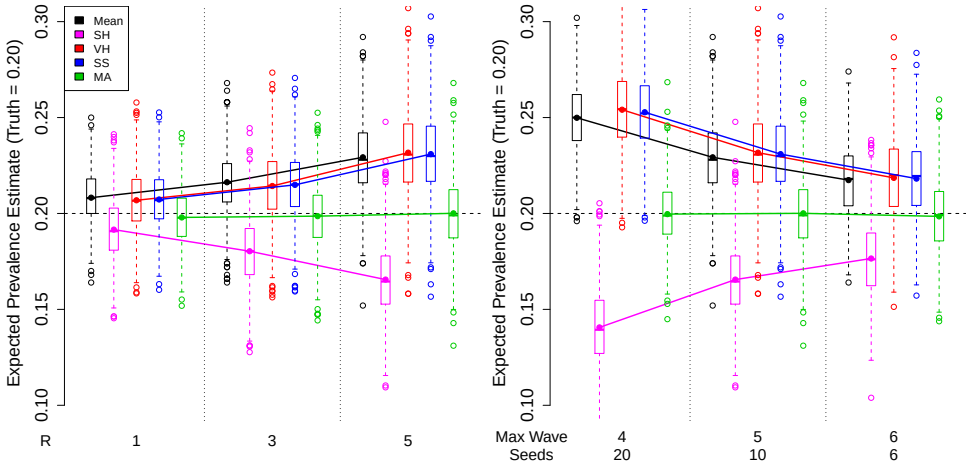
Figure 1: Comparison of performance of five RDS estimators under four conditions. $\hat{\mu}_{SS}$ and $\hat{\mu}_{MA}$ assume correct population size N. Results from 1000 simulations.

ately adjust for the over-sampling of infected nodes, resulting in unbiased estimators without increased variance.

The third section of Figure 1 considers the case the new estimator, $\hat{\mu}_{MA}$, is designed to address. There is homophily ($R = 5$), and all seeds are selected from among the infected nodes. This case treats a smaller sample fraction ($N = 1000$) and activity ratio 1 ($w = 1$). Here, all of the earlier estimators exhibit bias due to the selection of seeds (note the direction of bias is different for $\hat{\mu}_{SH}$), while the proposed estimator appropriately corrects for the selection of seeds.

The final case considers the joint effects of large sample fraction ($N = 715$), non-activity ratio ($w = 1.8$), homophily ($R = 5$), and biased selection of seeds (all infected). Here, the sample mean over-represents the higher-degree and initially sampled infected nodes. $\hat{\mu}_{\mathrm{VH}}$ exhibits a strong negative bias, similar to that in the second case. The two effects jointly cause tremendous bias in $\hat{\mu}_{SH}$. $\hat{\mu}_{SS}$ is affected by seed bias, al-

though not by the sample fraction. Here, again, the proposed estimator correctly adjusts for all of these effects. Although in this example the effects of sample fraction/activity ratio are of larger magnitude than those of seed bias/homophily, in practice the relative magnitudes of these will vary across data sets.



(a) (b)

Figure 2: Comparison of the performance of five RDS estimators with biased seed selection and various levels of homophily and numbers of sampling waves. All treat $N = 1000, w = 1$, and all seeds are selected from among infected nodes only. The first subfigure illustrates the exacerbating effect of homophily on seed bias. The second illustrates the ameliorating effect of increased sampling waves on seed bias.

Figures 2(a) and 2(b) illustrate additional features important to the impact of seed selection on the bias of the sample mean and earlier estimators. In particular, Figure 2(a) illustrates that the bias is exacerbated by increased homophily in the underlying population, and Figure 2(b) illustrates that bias is attenuated by having more sampling waves (attained by selecting fewer seeds for fixed sample size and branching). In each of these cases, the proposed estimator has negligible bias. Note that for very high levels of homophily ($R = 13$), the proposed estimator was found to exhibit positive bias, but of much smaller magnitude than that of the other estimators.

4.3 Sensitivity to the Population Size Estimate

In practice, the size of the hidden population, N, may not be known. We therefore conduct a sensitivity analysis illustrating the performance of the proposed estimator in the case of an inaccurate estimate $\hat{N}$ of N. We consider cases of $N-\frac{1}{2}(N-n) = \hat{N}_s < N$ and $N+\frac{1}{2}(N-n) = \hat{N}_l > N$. For each treatment of $\hat{N}$, we treat the four cases in Figure 1. $\hat{\mu}$ and $\hat{\mu}_{\mathrm{VH}}$ correspond to the extreme cases of $\hat{N} = n$ and $\hat{N} = \infty$, respectively, so are plotted for reference alongside $\hat{\mu}_{SS}$ and $\hat{\mu}_{MA}$ for each level of $\hat{N}$.

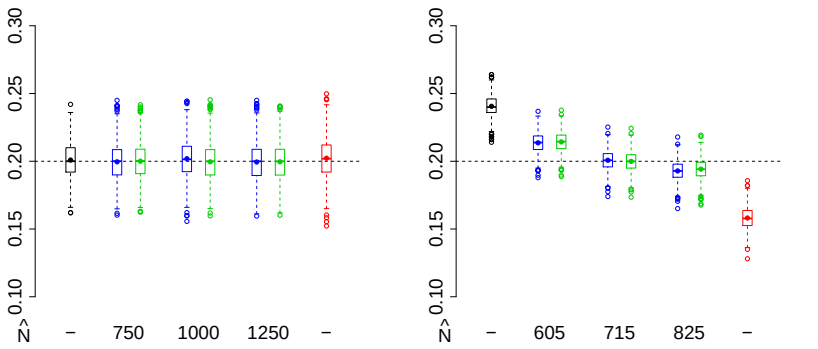
Figures 3(a) and 3(c) illustrate that in the case of unity activity ratio and a smaller sample fraction, the assumed population size has little impact on the resulting estimators (note that in the case with $\hat{N} = 715$ in Figure 3(c), seed bias leads to the perception of a non-unity activity ratio which, along with a smaller assumed population size results in the perception of a larger sample fraction).

Figures 3(b) and 3(d) illustrate that in the case of large sample fraction and activity ratio ($w \neq 1$), the assumed sample fraction does impact the estimators $\hat{\mu}_{SS}$ and $\hat{\mu}_{MA}$. When there is no seed bias (Figure 3(b)), these estimators perform nearly identically. Gile (2011) argues that in this case, $\hat{\mu}_{SS}$ interpolates between the sample mean and $\hat{\mu}_{\mathrm{VH}}$, and that trend seems to hold for $\hat{\mu}_{MA}$ as well. In the case of seed bias (Figure 3(d)), however, $\hat{\mu}_{SS}$ and $\hat{\mu}_{MA}$ differ, in that $\hat{\mu}_{MA}$ corrects for the bias induced by the seed selection.

Finally, it is worth noting that for smaller sample fractions, such as in Figure 3(c), the bias induced by seed selection may be of far greater magnitude than the bias induced in $\hat{\mu}_{\mathrm{VH}}$ by finite population effects. For this reason, for smaller sample fractions, $\hat{\mu}_{MA}$ may be able to correct for the more important form of bias, without being greatly affected by uncertainty in the population size.

4.4 Sensitivity to the network working model

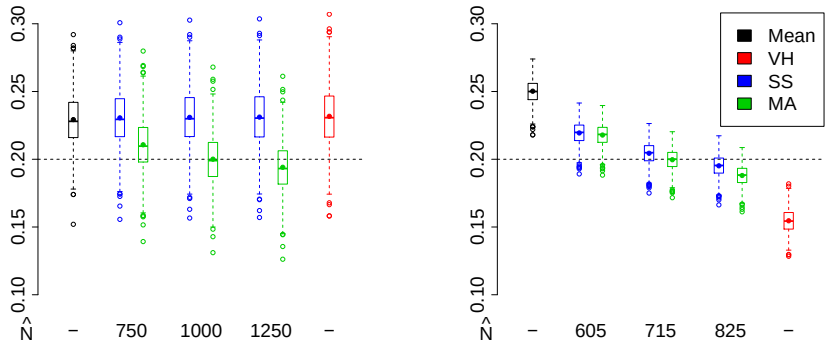
The role of the network working model is to provide a (stochastic) representation of the networked population. This model is the basis of the improved representation of the RDS design leveraged by the proposed estimator. The complexity of real-world social networks is high, so that simple network models will typically only capture a subset of this complexity.



(a) None (b) Sample Fraction, Activity Ratio

The ERGM in (2) is designed to represent two levels of network structure that are important to model RDS. The first is the nodal level representation of the individual heterogeneity in the propensity to have social ties. This is via the nodal degrees which are also a measure of the centrality of the individuals in the network (Freeman, 1979). The second is at the dyadic level and captures the homophily, or propensity for ties to be between individuals with the same infection status (beyond that implied by the infection prevalence). As infection status is the primary outcome of interest, this homophily is the most important to capture. The model (2) does not capture third level triadic effects, those based on the structure of triads of relations between individuals. While these are tertiary to the monadic and dyadic effects they can influence the RDS. Unfortunately RDS results in branching tree patterns of observations that limit the empirical information on these triadic effects. Hence the model (2) presumes that the triadic effects are those that would be produced by the modeled monadic and dyadic components.

The purpose of this section is to assess the sensitivity of the estimator to this misspecification of the triadic effects. Explicitly, we will consider networked populations with higher levels of transitivity than specified in the network working model and compare the performance of the estimators. Transitivity is represented by the edgewise shared partner (alter) statistics, denoted $\mathrm{EP}_0(\mathbf{y}), \ldots, \mathrm{EP}_{N-2}(\mathbf{y})$, where $\mathrm{EP}_k(\mathbf{y})$ is defined as the number of unordered pairs i, j such that $\mathbf{y}_{ij} = 1$ and i and j have exactly k common alters. It is a measure of the shared friendliness of friends. The geometri-



(c) Homophily, Seed Bias (d) All

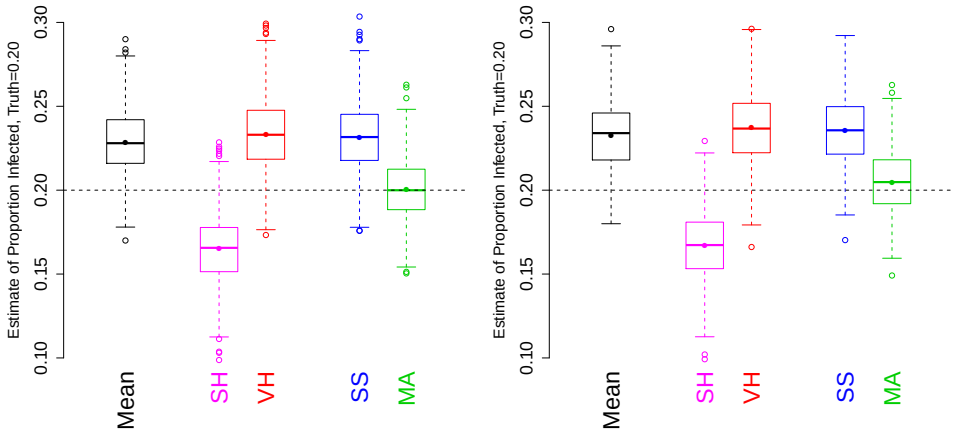
Figure 3: Sensitivity of $\hat{\mu}_{SS}$ and $\hat{\mu}_{MA}$ to inaccurate population size under four network and sampling conditions. Gile (2011) argues that $\hat{\mu}_{SS}$ approaches the sample mean for small assumed population sizes, and approaches $\hat{\mu}_{\text{VH}}$ for large assumed population sizes.

cally weighted edgewise shared partner (GWESP) statistic, conditional on the θ parameter, is

$$\text{gwesp}_\theta(\mathbf{y}) \equiv e^\theta \sum_{i=1}^{N-2} \left[1 - (1 - e^{-\theta})^i\right] \text{EP}_i(\mathbf{y}) \qquad \theta \geq 0.$$

The GWESP is an aggregate measure of local clustering or the overall "inwardness" of ties. The parameter θ controls just how "local" the clustering needs to be. If $\theta = 0$ an edge with one shared partner counts the same as an edge with two or more shared partners. If $\theta > 0$ an edge with one shared partner counts *less than* an edge with two or more shared partners. So large values of θ mean that very tight clustering is highly weighted and loose clustering is emphasized less. These terms have been developed for ERGM by Snijders, Pattison, Robins and Handcock (2006) and Hunter and Handcock (2006).

Most real-world networked populations over which RDS will be applied may be expected to have higher levels of transitivity than that produced by monadic and dyadic effects. Here we will consider two ways to produce networks with higher propensities for "friends of friends to be friends". To investigate the relative performance of the estimator in populations with higher transitivity, we generate networks with exactly the same monadic and dyadic statistics as those considered in Section 4.2 but



(a) Transitivity via GWESP with $\theta = 0$. (b) Transitivity via GWESP with $\theta = 1$.

Figure 4: Comparison of performance of five RDS estimators when the network working model misspecifies the transitivity. The populations in the left panel have ten times the GWESP with $\theta = 0$ and the right panel with $\theta = 1$. Results from 1000 simulations.

with higher transitivity as measured by the GWESP statistic. We do this by adding a $\text{gwesp}_\theta(\mathbf{y})$ term to the model (9) and inflating the mean value parameter of the $\text{gwesp}_\theta(\mathbf{y})$ while holding the mean value parameters of the other terms, as well as the degrees of each node, fixed. We then simulate networks from the resulting model, and apply the sampling and estimation procedures.

To test the correction for seed bias under model misspecification, we consider the populations in the third section of Figure 1. These have a smaller sample fraction ($N = 1000$), high homophily ($R = 5$), no differential activity ($w = 1$) and biased selection of seeds (all infected). We take the same 1000 populations and re-generate them with the exactly the same degree sequences and homophily but with increased transitivity (as measured by the $\text{gwesp}_\theta(\mathbf{y})$).

Figure 4 compares the same estimators as in Figure 1. The left panel consider populations with $\text{gwesp}_0(\mathbf{y})$ ten times that in the original. A value of $\theta = 0$ means that the statistic measures the number of pairs of people that are connected both by a direct edge *and* by a two-path through another person (that is, the number of edges minus the number of edges

connected by no two-paths). As can be seen, the performance of $\hat{\mu}_{MA}$ is little effected by the increased transitivity. The right panel compares populations with ten times $\text{gwesp}_1(y)$. A value of $\theta = 1$ means that the statistic weighs up the connectedness of edges with more weight on the terms with more shared partners. In this case $\hat{\mu}_{MA}$ has modest positive bias (0.46%) and similar variance compared to the estimators on the original populations.

5 Application to HIV prevalence in Hidden Populations

We apply our estimator to data collected in 2007 among injecting drug users (IDU) in Mykolaiv, Ukraine. The HIV epidemic in the Ukraine is one of the most severe in Europe, and still growing. As of 2009, the adult HIV prevalence was estimated at 0.86% (Ukrainian AIDS Centre, 2009). Ukraine's epidemic is most severe among injecting drug users and their sexual partners, who account for the majority of new infections (United States Agency for International Development, 2010). The data we consider here were collected as part of a series of studies of IDU across major Ukrainian cities in 2007 (Kruglov, Kobyshcha, Salyuk, Varetska, Shakarishvili and Saldanha, 2008). We focus on the data collected in Mykolaiv because, by chance and because of the contacts available to the researchers, all seeds in this sample were HIV positive.

This study began with 6 seeds and continued until wave 10, with 31 samples from wave 10, and a total of 260 samples. The average wave number was 6.1. The homophily based on HIV status for the population is estimated to be $R = 2.47$ and the differential activity is estimated to be 0.72.

Although the size of the population was not known precisely, an estimated range of population sizes is available through scale-up and multiplier methods (Kruglov et al., 2008; UNAIDS/WHO, 2003). We chose a population size, $N = 4000$, near the low end of this range. The variability of population size estimates is quite large, with a point estimate closer to 8000 in 2008 (Berleva, Dumchev, Kobyshcha, Paniotto, Petrenkon, Saliuk and I. A. Shvab, 2010)). We used sensitivity analysis to verify that population size 4000 is sufficiently large that our estimates are insensitive to increased population size.

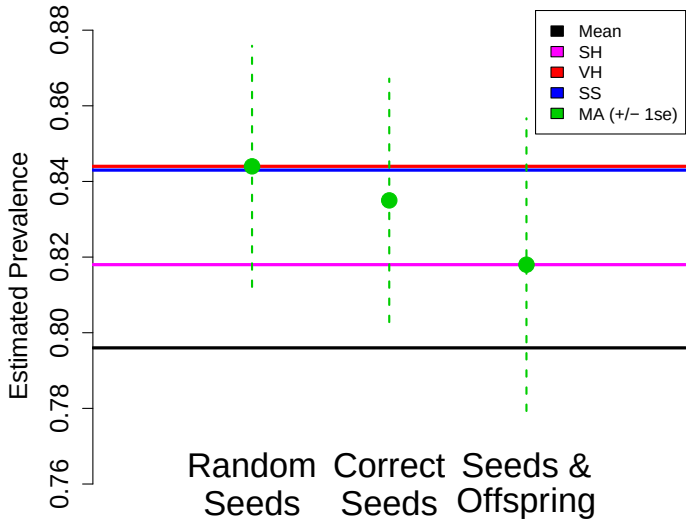


Figure 5: HIV prevalence estimates for injecting drug users in Mykolaiv, Ukraine, 2007.

We compare estimates for this application based on the current standard estimators ($\hat{\mu}_{SH}$, $\hat{\mu}_{\text{VH}}$, $\hat{\mu}_{SS}$) and three variants of $\hat{\mu}_{MA}$, summarized in Figure 5. First, we consider a version of $\hat{\mu}_{MA}$ which does not correct for seed bias. In this case, we select the seeds of the simulated samples with probability proportional to degree and without regard to infection status. As illustrated in Figure 5, this results in an estimate very close to that given by $\hat{\mu}_{SS}$ and $\hat{\mu}_{\text{VH}}$ ($\hat{\mu}_{\text{VH}} = 0.844$, $\hat{\mu}_{SS} = 0.843$, $\hat{\mu}_{MA} = 0.845$). In the second condition, we then apply the correction for seed bias, by matching the simulated seeds to the infection and degree characteristics of the observed seeds. This results in the second estimate in Figure 5, $\hat{\mu}_{MA} = 0.837$. This adjustment is in the direction we would expect, decreasing the prevalence estimate, corresponding to down-weighting the group over-represented in the seeds. The modest magnitude of this adjustment can be attributed to the weak homophily in this network, relative to the number of sample waves, as well as the high prevalence, leading to a smaller difference between prevalence in the population and prevalence among the seeds than in our simulation study.

The third condition we considered highlights the flexibility and possibilities for extensions of $\hat{\mu}_{MA}$. We note that in these data, infection groups

differed in their recruitment behavior. Some differences in recruitment behavior have been referred to as differential *recruitment effectiveness* in Heckathorn (2007), as well as in Tomas and Gile (2010) and Guntuboyina, Barbour and Heimer (in preparation). This is a pattern in which one group systematically recruits more effectively than another group. In this case, however, the pattern was more complex. On average, infected and uninfected participants did not vary greatly in their recruitment effectiveness. However, uninfected participants *in the early waves* of the study recruited disproportionately few additional participants, as illustrated in Figure 6. Because of the branching nature of the sampling, this resulted in a dramatic under-representation of uninfected IDU in the survey. To correct for this, however, we needed to estimate and replicate offspring distributions varying by both infection status and survey wave.

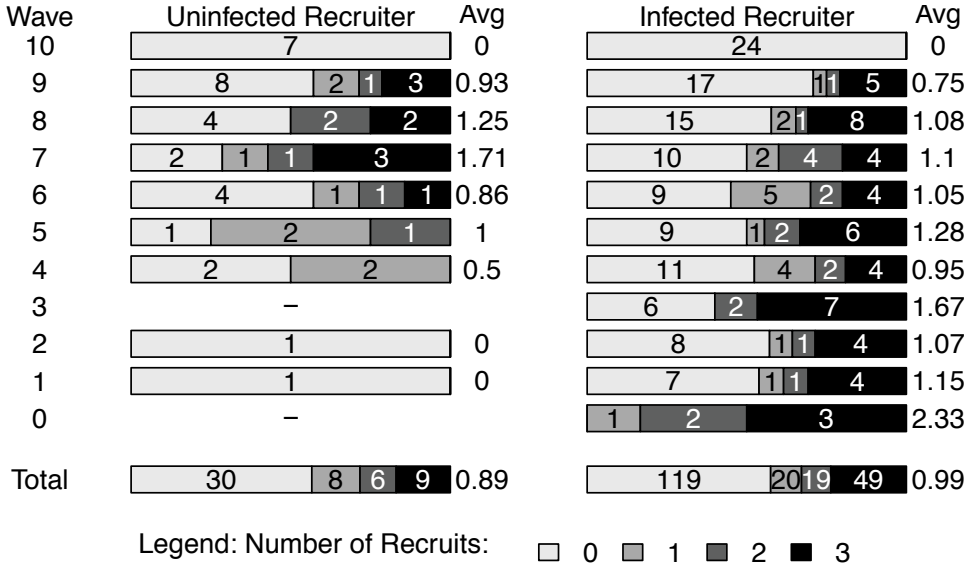


Figure 6: Distribution of number of successful recruitments by wave and infection status of recruiter. Uninfected recruiters were rare and unsuccessful in early waves, contributing to under-representation of uninfected participants in the sample. There were no uninfected participants in waves 0 (seeds) or 3.

We therefore applied a version of $\hat{\mu}_{MA}$, modified to reflect the empirical offspring distribution by wave and infection status. In most cases, this required simply assigning an offspring distribution equal to the empirical offspring distribution by wave and infection status. For uninfected re-

cruiters in wave 3, we used averaged empirical values from waves 2 and 4. For waves 10 and beyond, we replicated the empirical results from wave 9. Whenever a simulated recruiter did not have enough eligible alters to allow for the number of recruits selected from the appropriate distribution, we assigned any unfulfilled recruitments to the next active recruiters of the same infection status with fewer than 3 assigned recruits. This is straightforward to apply in our model-assisted setting and illustrates how this approach allows the approximation to the sampling design to be improved using available information.

The results of this analysis are illustrated in the third bar of Figure 5. The resulting estimate, 0.817, was substantially lower than the earlier estimates, suggesting the offspring distribution had a substantial impact on the resulting estimates.

6 Discussion

In this article we introduce a new approach to estimation based on RDS data that uses a working model for the underlying networked population to more accurately estimate the inclusion probabilities necessary for design-based inference.

We demonstrate that this approach allows us to correct for differential sampling probabilities based on nodal degrees, as in earlier RDS estimators (Heckathorn, 2007; Salganik and Heckathorn, 2004; Volz and Heckathorn, 2008), as well as for finite population biases, as addressed by another earlier approach (Gile, 2011). In addition, our proposed estimator is able to adjust for the convenience sample of seeds, a feature not accounted for in any previous approaches.

We apply this approach to obtain improved estimation of HIV prevalence in an IDU population in the Ukraine. We improve the approximation to the actual RDS process resulting in improved estimates, and compute associated measures of uncertainty. We also show the flexibility of the working model approach. It allows for additional information available in a particular application to be incorporated via the ERGM framework, and leverages recent advances in that area (Handcock, Hunter, Butts, Goodreau and Morris, 2008; Snijders et al., 2006). A significant weakness of our approach is the requirement that the population size is known. Our simulation study illustrates that the proposed estimator is indeed sensitive to estimates of population size, but as long as the population size is not

greatly under-estimated, we do not expect it to perform worse than the earlier estimator $\hat{\mu}_{\mathrm{VH}}$, and in cases of bias induced by the sample of seeds, $\hat{\mu}_{MA}$ may perform substantially better than any existing estimator, even for highly inaccurate assumptions regarding the population size.

We therefore propose three practical regimes for the use of $\hat{\mu}_{MA}$. First, if the population size is known, the estimator may be applied directly. If the population size is unknown, but a range of estimates is available, the estimator may be applied across the range. If the results vary greatly, the uncertainty of the resulting estimator should be adjusted accordingly. Such may be the case, for example, in populations of men who have sex with men, who are assumed to constitute 1% - 3% of many populations. Finally, if no information on population size is available, $\hat{\mu}_{SS}$ may be compared with $\hat{\mu}_{\mathrm{VH}}$ to determine whether there are important finite population effects in this sample. If not, $\hat{\mu}_{MA}$ may then be applied to correct for any effects of seed bias. In the case where finite population effects are found, $\hat{\mu}_{MA}$ may be applied to diagnose the extent of seed bias at each of several estimates of population size. Note that in such cases, the earlier estimators also make an assumption about population size (i.e. that it is *sufficiently large*), and so do not avoid the problem of unknown population size.

Another important assumption is the form of the social network working model. Our estimator relies on a simple model, not because we believe it to be strictly accurate, but because we expect it to capture the network features most important to the sampling process, and because it is feasible to estimate from the available data. To assess the sensitivity of the estimator to the form of the working model we considered versions of populations with greatly increased transitivity, a feature not captured by the working model. The results indicate only modest impact of high transitivity on the estimator or it uncertainty. While this may not be universally true, it does indicate the ability of the working models to capture nodal and dyadic effects goes a long way to improve the representation of the RDS process.

Several extensions of this approach are possible. First, if data on the characteristics of all alters are not available, we may wish to estimate the sum of cross-group ties ($g(\mathbf{y}, \mathbf{x})$) based on referral patterns. Such an estimate is used in the application to HIV prevalence estimation (Section 5).

Our approach can also be extended to include additional measurable features of the network working model or sampling process, such as ho-

mophily on neighborhood of residence or bias in the passing of coupons. We illustrate one such extension in Section 5, in which we observe an aberrant pattern of recruitment by infection status, and adapt the estimator to condition on this pattern. Note that the resulting estimate is very close to that given by $\hat{\mu}_{SH}$. This is consistent with results in Tomas and Gile (2010) indicating that $\hat{\mu}_{SH}$ is not as susceptible to bias induced by differential recruitment effectiveness as $\hat{\mu}_{\mathrm{VH}}$ or $\hat{\mu}_{SS}$.

Further extensions of this approach will make it possible to consider the joint estimation of population size and prevalence, or correlations between multiple nodal variables. We explore these features in ongoing work.

We intend to make code available for these procedures in the R package `RDS` on CRAN (Handcock, Gile and Neely, 2009; R Development Core Team, 2007).

Inferential and Computation: This supplement presents specifics of the estimation algorithms and our approach to standard error estimation (RDSMAsupplement.pdf)

References

Abdul-Quader, A. S., Heckathorn, D. D., McKnight, C., Bramson, H., Nemeth, C., Sabin, K., Gallagher, K., and Jarlais, D. C. D. (2006), "Effectiveness of Respondent-Driven Sampling for Recruiting Drug Users in New York City: Findings from a Pilot Study," *Journal of Urban Health*, 83, 459–476.

Berleva, G. O., Dumchev, K. V., Kobyshcha, Y. V., Paniotto, V. I., Petrenkon, T. V., Saliuk, T. O., and I. A. Shvab, I. A. (2010), Analytical Report based on sociological study results Estimation of the Size of Populations Most-at-Risk for HIV Infection in Ukraine in 2009,, Technical report, International HIV/AIDS Alliance in Ukraine.

Bernhardt, A., Heckathorn, D., Milkman, R., and Theodore, N. (2006), "Documenting Unregulated Work: A Survey of Workplace Violations in New York City," *The Future of Work*, . **URL:** `http://www.unprotectedworkers.org`

Felix-Medina, M. H., and Monjardin, P. E. (2006), "Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations: A Bayesian-Assisted Approach," *Survey Methodology*, 32, 187–195.

Felix-Medina, M. H., and Thompson, S. K. (2004), "Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations," *Journal of Official Statistics*, 20, 19–38.

Frank, O. (1971), *The Statistical Analysis of Networks*, London: Chapman and Hall.

Frank, O. (2005), "Network Sampling and Model Fitting," in *Models and Methods in Social Network Analysis*, eds. J. S. P. Carrington, and S. S. Wasserman, Cambridge: Cambridge University Press, pp. 31–56.

Frank, O., and Snijders, T. A. B. (1994), "Estimating the size of hidden populations using snowball sampling," *Journal of Official Statistics*, 10(1), 53–67.

Frank, O., and Strauss, D. (1986), "Markov Graphs," *Journal of the American Statistical Association*, 81(395), 832–842.

Freeman, L. C. (1979), "Centrality in Social Networks: Conceptual Clarification," *Social Networks*, 1(3), 223–258.

Freeman, L. C. (1996), "Some antecedents of social network analysis," *Connections*, 19, 39–42.

Gile, K. J. (2011), "Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation," *Journal of the American Statistical Association*, 106, 135–146.

Gile, K. J., and Handcock, M. S. (2010), "Respondent-Driven Sampling: An Assessment of Current Methodology," *Sociological Methodology*, 40, 285–327. **URL:** `http://arxiv.org/abs/0904.1855v1`

Goodman, L. A. (1961), "Snowball Sampling," *Annals of Mathematical Statistics*, 32, 148–170.

Guntuboyina, A., Barbour, R., and Heimer, R. (in preparation), "Bootstrap-based population mean estimators for Respondent Driven Sampling,".

Handcock, M. S., and Gile, K. J. (2010), "Modeling Social Networks from Sampled Data," *Annals of Applied Statistics*, 4(1), 5–25.

Handcock, M. S., Gile, K. J., and Neely, W. W. (2009), **RDS***: R Functions for Respondent-Driven Sampling*, Hard-to-Reach Population Methods Research Group `http://hpmrg.org/`, Seattle, WA. R package version 0.10. **URL:** `http://CRAN.R-project.org/package=RDS`

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2008), "**statnet**: Software tools for the representation, visualization, analysis and simulation of social network data," *Journal of Statistical Software*, 24(1). **URL:** `http://www.jstatsoft.org/v24/i01/`

Heckathorn, D. D. (1997), "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations," *Social Problems*, 44, 174–199.

Heckathorn, D. D. (2007), "Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment," *Sociological Methodology*, 37, 151–207.

Heckathorn, D. D., and Jeffri, J. (2001), "Finding the beat: Using respondent-driven sampling to study jazz musicians," *Poetics*, 28, 307–329.

Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008), "Goodness of Fit for Social Network Models," *Journal of the American Statistical Association*, 103, 248–258.

Hunter, D. R., and Handcock, M. S. (2006), "Inference in curved exponential family models for networks," *Journal of Computational and Graphical Statistics*, 15(3), 565–583.

Johnston, L. G. (2010), "Starting RDS Session III: Seeds,". **URL:** `http://www.lisagjohnston.com/respondent-driven-sampling`

Kruglov, Y. V., Kobyshcha, Y. V., Salyuk, T., Varetska, O., Shakarishvili, A., and Saldanha, V. P. (2008), "The most severe HIV epidemic in Europe: Ukraine's national HIV prevalence estimates for 2007," *Sexually Transmitted Infections*, 84(Suppl 1), i37–41.

Lansky, A., Abdul-Quader, A. S., Cribbin, M., Hall, T., Finlayson, T. J., Garfein, R. S., Lin, L. S., and Sullivan, P. S. (2007), Developing an HIV behavioral surveillance system for injecting drug users: the National HIV Behavioral Surveillance System,, Technical Report Public Health Reports 2007; 122 Suppl 1: 48-55 17354527, Division of HIV/AIDS Prevention, National Center for HIV, STD, and TB Prevention, Centers for Disease Control and Prevention.

Lazarsfeld, P., and Merton, R. (1954), "Friendship as social process: A substantive and methodological analysis," in *Freedom and Control in Modern Society*, eds. M. Berger, T. Abel, and C. H. Page, New York: Van Nostrand, pp. 18–66.

Lee, T. C. M., and Meng, X.-L. (2007), "Self-Consistency: A General Recipe for Wavelet Estimation With Irregularly-spaced and/or Incomplete Data,". ArXiv Preprint. **URL:** `http://arxiv.org/abs/math/0701196v1`

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis with Missing Data, 2nd. ed.*, Hoboken, New Jersey: John Wiley & Sons, Inc.

Lumley, T. S. (2010), *Complex Surveys: A Guide to Analysis Using R*, New York: Wiley. Wiley Series in Survey Methodology.

Malekinejad, M., Johnston, L., Kendall, C., Kerr, L., Rifkin, M., and Rutherford, G. (2008), "Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance in International Settings: A Systematic Review," *AIDS and Behavior*, 12, 105–130.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001), "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, 27, 415–444.

Pattison, P., Robins, G., Daraganova, G., Wang, P., Koskinen, J., and Snijders, T. (2009), "Modelling large social networks: statistical issues,". Workshop Statistical Network Modeling, Nuffield College, Oxford.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, Version 2.6.1. **URL:** `http://www.R-project.org/`

Rubin, D. B. (1976), "Inference and missing data," *Biometrika*, 63, 581–592.

Salganik, M. J., and Heckathorn, D. D. (2004), "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling," *Sociological Methodology*, 34, 193–239.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Snijders, T. A. B., Pattison, P., Robins, G. L., and Handcock, M. S. (2006), "New Specifications for Exponential Random Graph Models," *Sociological Methodology*, 36, 99–153.

Thompson, S. K. (1990), "Adaptive Cluster Sampling," *Journal of the American Statistical Association*, 85, 1050–1059.

Thompson, S. K. (1992), *Sampling*, New York: Wiley.

Thompson, S. K. (2006), "Adaptive web sampling," *Biometrics*, 62(4), 1224–1234.

Thompson, S. K., and Frank, O. (2000), "Model-based estimation with link-tracing sampling designs," *Survey Methodology*, 26, 87–98.

Thompson, S. K., and Seber, G. A. F. (1996), *Adaptive sampling*, New York: Wiley.

Tomas, A., and Gile, K. J. (2010), "The Effect of Differential Recruitment, Non-response and Non-recruitment on Estimators for Respondent-Driven Sampling,". ArXiv Preprint. **URL:** `http://arxiv.org/abs/1012.4122`

Trow, M. (1957), *Right-Wing Radicalism and Political Intolerance*, New York: Arno Press. Reprinted 1980.

Ukrainian AIDS Centre (2009), National Estimate of HIV/AIDS Situation in Ukraine as of Beginning of 2009,, Technical report, Ministry of Health of Ukraine.

UNAIDS (2008), 2008 Report on the Global AIDS Epidemic,, Technical report, UNAIDS - Joint United Nations Programme on HIV/AIDS. **URL:** `http://www.unaids.org`

UNAIDS/WHO (2003), Estimating the size of populations at risk for HIV: Issues and Methods,, Technical report, UNAIDS/WHO Working Group on HIV/AIDS. **URL:** `http://www.unaids.org`

United States Agency for International Development (2010), HIV/AIDS Health Profile, October 2010,, Technical report, USAID/Ukraine. **URL:** `http://ukraine.usaid.gov/`

Volz, E., and Heckathorn, D. D. (2008), "Probability Based Estimation Theory for Respondent Driven Sampling," *Journal of Official Statistics*, 24(1), 79–97.

Watters, J. K., and Biernacki, P. (1989), "Targeted Sampling: Options for the Study of Hidden Populations," *Social Problems*, 36(4), 416–430.

Supplemental Materials: Network Model-Assisted Inference from Respondent-Driven Sampling Data

August 2, 2011

1 Estimation Procedure

We propose the following algorithm to compute the new estimator $\hat{\mu}_{MA}$ of μ.

1. Estimate the following according to their empirically observed values:
   - Sample size n
   - Number of seeds n^{seeds}, and degree and infection status of seeds, given by $\mathbb{N}^{seeds} = \{\mathbb{N}_{ij}^{seeds}\}$, where $\mathbb{N}_{ij}^{seeds}$ represents the number of seeds with degree i and infection j, $i \in 1 \ldots N-1$, $j \in \{0,1\}$.
   - Offspring distributions $\mathbf{p}^s$, where $\mathbf{p}_i^s$ = the proportion of the sample with i offspring, $i = 0, 1, \ldots,$ maximum number of coupons.
2. Estimate:
$$\hat{\pi}_i^0 = \frac{d_i}{N} \sum_{j=1}^{N} \frac{S_j}{d_j}, \quad i : S_i = 1.$$
3. For $r = 1 \ldots h$:

(a) Estimate:

$$\tilde{\mathbb{N}}^r_{kl} = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{S}_i \mathbb{I}(\mathbf{d}_i = k, \mathbf{z}_i = l)}{\hat{\pi}_i^{r-1}}$$

$$\tilde{g}(\mathbf{y}, \mathbf{x})^r = \sum_{i=1}^{N} \frac{\mathbf{S}_i \left(\mathbf{x}_i(1 - \mathbf{z}_i) + (\mathbf{d}_i - \mathbf{x}_i)\mathbf{z}_i\right)}{2\hat{\pi}_i^{r-1}}$$

(b) Compute the ERGM parameter η in the model (2) based on $\tilde{\mathbb{N}}^r$ and $\tilde{g}(\mathbf{y}, \mathbf{x})^r$ via the procedure in Supplemental Section 3. Denote the estimate by η^r. This step is conducted using the `statnet` R package (Handcock, Hunter, Butts, Goodreau and Morris, 2003).

(c) Simulate M_1 networks according to the distribution given by $\hat{\eta}^r, \tilde{\mathbb{N}}^r$, and $\tilde{g}(\mathbf{y}, \mathbf{x})^r$, also using the `statnet` R package.

(d) Simulate M_2 RDS samples from each of the M_1 networks in the previous step, according to sampling parameter $\mathscr{S} = \{n, \mathbb{N}^{seeds}\}, \mathbf{p}^s$. Let U^r_{kl} represent the number of times a node of degree k and infection l is sampled, over all $M = M_1 \times M_2$ samples.

(e) Estimate $\hat{\pi}^r_i \; \forall \; i : \mathbf{S}_i = 1$ in a manner similar to Fattorini (2006) and Gile (2011):

$$\hat{\pi}^r_i = \frac{U^r_{\mathbf{d}_i\mathbf{z}_i} + 1}{M \cdot \mathbb{N}^r_{d_i z_i} + 1}$$

4. Let $\hat{\pi}_i = \hat{\pi}^h_i$

5. Estimate

$$\hat{\mu}_{MA} = \frac{\sum_{i=1}^{N} \frac{\mathbf{S}_i \mathbf{z}_i}{\hat{\pi}_i}}{\sum_{i=1}^{N} \frac{\mathbf{S}_i}{\hat{\pi}_i}} \tag{S.1}$$

The simulations in this paper are based on $r = 3$ iterations, each including $M_1 = 25$ network samples and $M_2 = 20$ RDS samples from each network. In general, we recommend at least $M_1 = 25$, $M_2 = 20$ and $r = 3$. Estimation time scales with sample size, population size, and M. In our simulations, with $N = 1000$, estimates require about 20 minutes on a personal computer. In practice, these parameters can be adjusted for desired precision in the solution to (7). We will make available the code to compute this estimator in the `RDS` R package on CRAN (Handcock, Gile and Neely, 2009; R Development Core Team, 2007).

2 Measures of Uncertainty: Bootstrap

Unlike earlier RDS estimators, the estimator given in (7) allows for estimators of uncertainty that account for the estimated full relational structure of the underlying population as well as incorporating several observable features of the sampling process. The former is because of the use of the network working model for the population over which the RDS sampling procedure operates. The latter is because our procedure enables the simulation of complex RDS designs. In particular, if we believe there is seed bias we can incorporate it into the sampler, and if there is measurable sampling bias (as in the application) we can incorporate that also. This allows the procedure to incorporate available information about the population and sampling and greatly improves the accuracy of the representation of the actual sampling process. The accuracy of the bootstrap depends directly on the quality of the approximation to the actual sampling process.

We propose a parametric bootstrap approach to obtaining confidence intervals, according to the following procedure:

1. For $b = 1, \ldots B$, iterate the following steps:
   (a) Simulate a network $\mathbf{Y}_b$ from the model given in (2) with parameters $\eta = \hat{\eta}^h$, and $\mathbb{N}^h$ where h is the final iteration of the algorithm in Supplemental Section 1.
   (b) Simulate one RDS sample $S_{\mathbf{Y}_b}$ with parameter $\mathscr{S}_b$ from $\mathbf{Y}_b$.
   (c) Compute an estimator $\hat{\mu}_{MA}(b)$ of μ based on the sample $S_{\mathbf{Y}_b}$ using the algorithm in Supplemental Section 1.
2. Use the empirical distribution $\hat{\mathscr{D}}(\hat{\mu}_{MA})$ of $\{\hat{\mu}_{MA}(1), \hat{\mu}_{MA}(2), \ldots \hat{\mu}_{MA}(B)\}$ to estimate the distribution of $\hat{\mu}_{MA}$ under the estimated model form.

The distribution of $\hat{\mathscr{D}}(\hat{\mu}_{MA})$ may then be used to form confidence intervals for μ which account for the full estimated relational structure as well as observable biases of the sampling process. We use the standard deviation of the resulting population of B bootstrap estimates as an estimate of the standard error of $\hat{\mu}_{MA}$. We have used $B = 1000$ bootstrapped samples. In our simulations, this procedure took about 20 minutes per sample on a single processor. Parallelization is straightforward and dramatically reduces elapsed time. A large additional speedup can be obtained by replacing step (c) with (c′) in which the weight *classes* $\hat{\pi}_i^h$ from (S.1) are reused to

weight each bootstrap sample. While these weights vary from bootstrap sample to sample, their uncertainty is a small part of the overall uncertainty. This reduces the procedure to about 30 seconds per sample on a single processor. The analysis below uses (c').

We illustrate the performance of this standard error estimator by comparing five critical cases. As with the point estimate, we illustrate both cases in which we expect the estimator to perform reasonably well, and a case in which we expect the estimator to perform poorly. We introduce various forms of sampling biases. The initial sample can be selected either independent of infection status (denoted "No" in the bias column of Table 1) or all from within the infected subgroup ("Initial" bias). We also introduce referral bias where all infected alters are 20% more likely to be referred than uninfected alters ("Referral" bias).

Each set of simulations involved 1000 bootstrapped re-samples for each of 1000 simulated RDS samples. The parameters of the samples, average estimated standard errors, and coverage rates of nominal 95% and 90% confidence intervals are given in Table 1.

Table 1: Observed (simulation) standard errors of estimates, and average bootstrap standard error estimates, along with coverage rates of nominal 95% and 90% confidence intervals for procedure given in Supplemental Section 2 for varying sample proportion, homophily R, and activity ratio w, and for various biases in the sample selection process. Observed standard errors are based on 1000 samples. Bootstrap standard errors are the average bootstrap standard error estimates over the same 1000 samples. Nominal confidence intervals are based on quantiles of the Gaussian distribution.

% sample	homoph. R	w	sample bias	SE observed	SE bootstrap	coverage 95%	coverage 90%
50%	1	1	No	0.0140	0.0137	94.1%	88.8%
70%	1	1.8	No	0.0073	0.0075	94.9%	90.4%
50%	5	1	Initial	0.0188	0.0175	93.7%	87.9%
50%	5	1.8	Initial	0.0079	0.0080	95.0%	87.3%
50%	5	1	Referral	0.0216	0.0225	91.7%	84.7%

The magnitudes of the average bootstrap standard error estimates are quite close to the observed values in the first four cases, and the coverage rates in the cases without referral bias are very close to their nominal values. In this last case, the standard error estimator is anti-conservative because the bootstrap procedure does not replicate the referral bias in the sample.

The last row of Table 1 illustrates the poor performance of the estimator in the case of extreme referral bias. In this case, the estimator $\hat{\mu}_{MA}$ has positive bias (1.74%), leading to moderately lower coverage rates of the nominal intervals.

3 Inference for the ERGM conditional on the degree and infection status sequences

The model-assisted approach is based on a "working" model (2) for the networked population. The unknowns in the model are the finite-population values $\mathbf{d}$ and $\mathbf{z}$ and the super-population parameter η. Finite-population estimates of $\mathbb{N}$ (i.e., $\mathbf{d}$ and $\mathbf{z}$) and $g(\mathbf{y}, \mathbf{x})$ are determined by design-based inference as in (5) and (6). The estimate of η is computed as the natural parameter in (2) corresponding to these values. That is, the natural parameter corresponding to the mean-value parameter $\tilde{g}(\mathbf{y}, \mathbf{x})$ conditional on the degree sequence $\mathbf{d}$ and infection status sequence $\mathbf{z}$ induced by $\tilde{\mathbb{N}}$. Explicitly, we construct the joint degree and infection status sequence $\mathbf{d}, \mathbf{z}$ from $\mathbb{N}$, where the ordering of nodes is arbitrarily assigned (w.l.o.g.). To compute η we construct a network with this joint degree and degree status sequence and cross-group contacts $g(\mathbf{y}, \mathbf{z})$ using the Reed-Molloy method and then simulated annealing (Handcock et al., 2003; Handcock, Hunter, Butts, Goodreau and Morris, 2008; Molloy and Reed, 1995).

We can then compute $\hat{\eta}$ using the Geyer-Thompson MCMC approach (Handcock et al., 2003, 2008). As this is computationally expensive and unstable in this situation we use an approach based on a form of pseudo-likelihood introduced below.

Consider a model similar to (2) but with network space $\mathscr{Y}$ consisting of all binary undirected networks (i.e., unconditional on $\mathbf{d}$ and $\mathbf{z}$). Until recently inference for such models have been almost exclusively based on a local alternative to the likelihood function referred to as the *pseudo-*

likelihood (Besag, 1975; Strauss and Ikeda, 1990). Consider the conditional formulation of this model:

$$\text{logit}[P(Y_{ij} = 1 | Y_{ij}^c = \mathbf{y}_{ij}^c, \eta)] = \eta\delta(\mathbf{y}_{ij}^c) \qquad \mathbf{y} \in \mathscr{Y} \tag{S.2}$$

where $\delta(\mathbf{y}_{ij}^c, \mathbf{z}) = g(\mathbf{y}_{ij}^+, \mathbf{z}) - g(\mathbf{y}_{ij}^-, \mathbf{z})$, the change in $g(\mathbf{y}, \mathbf{z})$ when $\mathbf{y}_{ij}$ changes from 0 to 1 while the remainder of the network remains $\mathbf{y}_{ij}^c$ (See Strauss and Ikeda, 1990). The pseudo-likelihood for the model is:

$$\ell_P(\eta; \mathbf{y}) \equiv \eta \sum_{ij} \delta(\mathbf{y}_{ij}^c, \mathbf{z}) \mathbf{y}_{ij} - \sum_{ij} \log\left[1 + \exp(\eta\delta(\mathbf{y}_{ij}^c, \mathbf{z}))\right]. \tag{S.3}$$

This is the standard form of pseudo-likelihood, which we refer to a dyadic pseudo-likelihood.

This form is algebraically identical to the likelihood for a logistic regression model where each unique element of the adjacency matrix, $\mathbf{y}_{ij}$, is treated as an independent observation with the corresponding row of the design matrix given by $\delta(\mathbf{y}_{ij}^c, \mathbf{z})$. Then the maximum likelihood estimate (MLE) for this logistic regression model is identical to the maximum dyadic pseudo-likelihood (MPLE) for the corresponding ERG model, a fact that is exploited in computation. Therefore, algorithms to compute the MPLE for ERGMs are typically deterministic while the algorithms to compute their MLEs are typically stochastic. In addition, algorithms to compute the MLE can be unstable if the model is near degenerate (Handcock, 2003). This can lead to computational failure.

This standard form of pseudo-likelihood is inappropriate for the model (2) as it does not take into account the network space $\mathscr{Y}(\mathbf{z}, \mathbf{d})$. This is because $P(\mathbf{Y}_{ij} = 1 | \mathbf{Y}_{ij}^c = \mathbf{y}_{ij}^c, \eta)$ is either 1 or 0 depending on if the value of $\mathbf{y}_{ij}$ because the model conditions on the degree sequence consistent with $\mathbf{d}$. Hence the MPLE will usually produce non-sensical results.

Instead of a dyadic pseudo-likelihood we develop a tetradic pseudo-likelihood. We focus on ordered dyad-quads $\mathbf{y}_{ijkl} = (\mathbf{y}_{ij}, \mathbf{y}_{kl}, \mathbf{y}_{il}, \mathbf{y}_{jk})$ such that $\mathbf{y}_{ij} = \mathbf{y}_{kl} = 1, \mathbf{y}_{il} = \mathbf{y}_{jk} = 0$. We refer to this configuration as $\mathbf{y}_{ijkl}^+$. For each such dyad-quad there exists an alternative realization in which $\mathbf{y}_{ij} = \mathbf{y}_{kl} = 0, \mathbf{y}_{il} = \mathbf{y}_{jk} = 1$. We refer to this configuration as $\mathbf{y}_{ijkl}^-$. Thus $\mathbf{y}_{ijkl}^+$ and $\mathbf{y}_{ijkl}^-$ represent a pair in which $\mathbf{y}_{ij}$ is toggled from 1 to 0 in such a way as to retain the degree and infection status sequences of the corresponding full network.

Let $\mathbf{Y}_{ijkl} = (\mathbf{Y}_{ij}, \mathbf{Y}_{kl}, \mathbf{Y}_{il}, \mathbf{Y}_{jk})$, $\mathbf{Y}^c_{ijkl} = \mathbf{Y} \backslash \mathbf{Y}_{ijkl}$ and $\mathbf{y}^c_{ijkl} = \mathbf{y} \backslash \mathbf{y}_{ijkl}$. Let $\mathscr{D} = \{ijkl : \mathbf{y}^c_{ijkl} \cup \mathbf{y}^+_{ijkl} \in \mathscr{Y}(\mathbf{z}, \mathbf{d})\}$. For these dyad-quad configurations we then have:

$$\text{logit}[P(\mathbf{Y}_{ijkl} = \mathbf{y}^+_{ijkl} | \mathbf{Y}^c_{ijkl} = \mathbf{y}^c_{ijkl}, \eta)] = \eta \delta(\mathbf{y}^c_{ijkl}, \mathbf{z}) \qquad ijkl \in \mathscr{D} \quad \text{(S.4)}$$

where $\delta(\mathbf{y}^c_{ijkl}, \mathbf{z}) = g(\mathbf{y}^c_{ijkl} \cup \mathbf{y}^+_{ijkl}, \mathbf{z}) - g(\mathbf{y}^c_{ijkl} \cup \mathbf{y}^-_{ijkl}, \mathbf{z})$, the change in $g(\mathbf{y}, \mathbf{z})$ when $\mathbf{y}_{ijkl}$ changes from $\mathbf{y}^-_{ijkl}$ to $\mathbf{y}^+_{ijkl}$. The tetradic pseudo-likelihood for model (2) can then be defined as:

$$\ell_{PT}(\eta; \mathbf{y}) \equiv \eta \sum_{ijkl \in \mathscr{D}} \delta(\mathbf{y}^c_{ijkl}, \mathbf{z}) \mathbb{I}(\mathbf{y}_{ijkl} = \mathbf{y}^+_{ijkl}) - \sum_{ijkl \in \mathscr{D}} \log \left[1 + \exp(\eta \delta(\mathbf{y}^c_{ijkl}, \mathbf{z}))\right]. \quad \text{(S.5)}$$

As $|\mathscr{D}|$ is large, we take a large random sample of them ($N = 100000$) and use the sample mean to approximate (S.5). This procedure is implemented in the `statnet` `R` package (Handcock et al., 2003).

While the MPLE is know to be inferior to the MLE for dyadic dependence models (van Duijn, Handcock and Gile, 2009) it is equivalent to the MLE for some dyadic independence models. For the model (2) the network statistic is weakly dependent on the set of networks with the given degree and infection sequences. Hence the maximum tetradic pseudo-likelihood (MTPLE) might be expected to perform well for this model. This does seem to be the case for the models considered in this paper. In simulations (not shown here) as it appears to be indistinguishable from the MLE (where the later is computed by a computationally expensive MCMC procedure). The advantages of the tetradic MPLE are that it is computationally stable and fast while being numerically indistinguishable from the MCMC-MLE. For these reasons we use it in all simulations in this paper.

This estimator could be improved by adding hexadic configurations to the pseudo-likelihood. These are necessary for sampling algorithms to cover the full network space (Rao, Jana and Bandyopadhyay, 1996). However they also lead to more complex algorithms and will be considered in other work.

References

Besag, J. (1975), "Statistical analysis of non-lattice data," *The Statistician*, 24, 179–95.

Fattorini, L. (2006), "Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities," *Biometrika*, 93(2), 269–278.

Gile, K. J. (2011), "Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation," *Journal of the American Statistical Association*, 106, 135–146.

Handcock, M. S. (2003), Assessing Degeneracy in Statistical Models of Social Networks,, CSSS working paper no 39, University of Washington.

Handcock, M. S., Gile, K. J., and Neely, W. W. (2009), **RDS***: R Functions for Respondent-Driven Sampling*, Hard-to-Reach Population Methods Research Group `http://hpmrg.org/`, Seattle, WA. R package version 0.10. **URL:** *`http://CRAN.R-project.org/package=RDS`*

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2003), **statnet***: Software Tools for the Statistical Modeling of Network Data*, Statnet Project `http://statnet.org/`, Seattle, WA. R package version 2.0. **URL:** *`http://CRAN.R-project.org/package=statnet`*

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2008), "**statnet**: Software tools for the representation, visualization, analysis and simulation of social network data," *Journal of Statistical Software*, 24(1). **URL:** *`http://www.jstatsoft.org/v24/i01/`*

Molloy, M. S., and Reed, B. A. (1995), "A critical point for random graphs with a given degree sequence," *Random Structures and Algorithms*, 6, 161–179.

Rao, A. R., Jana, R., and Bandyopadhyay, S. (1996), "A Markov chain Monte Carlo method for generating random $(0, 1)$ matrices with given marginals," *Sankhya, Series A*, 58, 225–242.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, Version 2.6.1. **URL:** *`http://www.R-project.org/`*

Strauss, D., and Ikeda, M. (1990), "Pseudolikelihood estimation for social networks," *Journal of the American Statistical Association*, 85, 204–212.

van Duijn, M. A. J., Handcock, M. S., and Gile, K. J. (2009), "A Framework for the Comparison of Maximum Pseudo Likelihood and Maximum Likelihood Estimation of Exponential Family Random Graph Models," *Social Networks*, 31, 52–62.